Towards Measuring the Complexity of Introducing Semantics into a Company

Liliana Ibeth Barbosa Santillán and Inmaculada Álvarez de Mon y Rego

Abstract— **The Semantics Difficulty Model (SDM) is a model that measures the difficulto of introducing semantics technology into a company. SDM manages three descriptions of stages, which we will refer to as "snapshots": a company semantic snapshot, data snapshot and semantic application snapshot.**

Understanding a priory the complexity of introducing semantics into a company is important because it allows the organization to take early decisions, thus saving time and money, mitigating risks and improving innovation, time to market and productivity.

SDM works by measuring the distance between each initial snapshot and its reference models (the company semantic snapshots reference model, data snapshots reference model, and the semantic application snapshots reference model) with Euclidian distances. The difficulty level will be "not at all difficult" when the distance is small, and becomes "extremely difficult" when the the distance is large.

SDM has been tested experimentally with 2000 simulated companies with arrangements and several initial stages. The output is measured by five linguistic values: "not at all difficult, slightly difficult, averagely difficult, very difficult and extremely difficult".

As the preliminary results of our SDM simulation model indicate, transforming a search application into integrated data from different sources with semantics is a "slightly difficult", in contrast with data and opinion extraction applications for which it is "very difficult".

I. Introduction

Companies have formal, semi-formal and informal knowledge that can be accessed, shared and reused by its members in order to solve their individual or collective tasks. This knowledge is structured by its semantics. Semantics refers to meaning. Here it is mainly a reference to a particular thing or notion. Therefore, to provide semantics, it is necessary to give an interpretation specifying the meaning of each symbol in a sentence of the language [Alonso-Ovalle(2006)]. Semantics could be static or dynamic, and implicit or explicit through the company. The semantics can be tangible or intangible (implicit or tacit) following the well-known distinction by Nonaka and Takeuchi [Nonaka and Takeuchi(1995)] and Uschold[Uschold and Gruninger(1996)].

For companies, explicit semantics is useful in connecting people to people, people to information, and information to information.

So far there has been little discussion about how to identify implicit semantics in order to measure the complexity of making it explicit. However, nowadays one of the most significant problems being discussed is that semantics needs to be understood by the electronic devices.

Today, semantic technologies are much more mature and solid. A lot of companies want to know the complexity of introducing semantics a priori so that they can make early decisions. Thus they can save time and money, mitigate risks, and improve innovation, quality, cost-effectiveness, time-to-market and productivity.

Ontologies [Gruber(1993)] have shown excellent results for modeling and structuring problems by providing a formal conceptualization of a particular domain that is shared by a group of people in a company. However, a major problem with this kind of explicit semantics is who is going to carry out this job? So we identified two types of companies. The first is a company that wants to introduce semantics and the second is a company in charge of introducing semantics to a tangible resource.

The aim of this paper is to start the study of the means to determine how difficult it is to introduce semantics into a company. We are motivated by the notion that *we can not predict what we can not measure.*

Recently, members of companies [Ontoprise(2012)], [IBM(2012)], [Oracle(2012)], [AG(2012)], [HP(2012)] have shown an increased interest in developing tools, solutions, and products to introduce semantic (web) applications. On the other hand, a considerable amount of literature has been published on effort estimation. These studies have shown interesting results. The first serious discussions and analysis of estimation emerged during the 1980s with the Constructive Cost Model (COCOMO [Center(1997)]) work. COCOMO is a screen-oriented, interactive software package that assists in budget planning and schedule estimation of a software project prior to beginning any work. With COCOMO [Center(1997)], a software project manager (or team leader) can develop a model (or multiple models) of a project in order to identify potential problems in resources, personnel, budgets, and schedules both before and after the software life cycle.

The second serious discussion arose in the software engineering field with Personal Software Process (PSP) [Humphrey(2005)]. PSP is a set of practices and methods that enables software developers to control their own working lives. When competent professionals learn and consistently follow these engineering and scientific principles, they are empowered to manage their own work and do an excellent job.

In recent years, there has been also an increasing amount of literature on effort estimation with Ontology Cost Model

*This work was supported by Universidad Politécnica de Madrid (UPM)

Facultad de Informática. UPM -Departamento de Matemática Aplicada - Madrid, Spain lbarbosa@alumnos.fi.upm.es

Ingenieria Técnica de Telecomunicación. UPM - Lingüistica aplicada a la ciencia y la tecnología- Madrid, Spain ialvarez@euitt.upm.es

(ONTOCOM)[Bontas and Tempich(2005)]. ONTOCOM is a cost estimation model for the area of Ontology Engineering. The goal of this model is to predict the costs arising in typical classes of ontology engineering processes such as ontology building, reuse or maintenance.

Perhaps the most serious disadvantage of these models and processes is that they focus on the development of both software and ontologies. Difficulties arise, however, when a company wants to measure a priori how difficult it is to introduce semantics. Nevertheless, the strategy used in COCOMO[Center(1997)], PSP[Humphrey(2005)] and ONTOCOM[Bontas and Tempich(2005)] is a useful starting point.

Our contribution is to measure the degree of difficulty involved in introducing semantic technology with different degrees of complexity.

The remainder of this paper is structured as follows. In section 2 the background and the model are described together with the assumptions and the limitations of our work. In section 3 our idea is presented and explained. In section 4 an analysis of the model is carried out. In section 5 the experimental results are shown and finally section 6 concludes.

II. Background and Model

We identify three dimensions: a company semantic snapshot, a data snapshot and a semantic application snapshot. Each dimension will be defined as follows:

Company semantic snapshot In this dimension, the semantics is visible and permanent by the process treatment, which is the procedure explained in section III.F.

Data snapshot In this dimension, the focus is onto the structured data and unstructured data in the company.

Semantic Application snapshot In this dimension, the focus is on the know-how on semantic infrastructure of the engineering department of an Information Tecnologies applications development company.

The first dimension will answer the following questions: how can the semantic level of the company be known? Which and how many semantic levels are there?

The second dimension will answer the following questions: what data content will be used by the company? What is the size of the data?

The last dimension will answer the following questions: which and how many types of semantic applications are there? What kind of application does a company want to introduce semantics into? What are the main metrics in the process to introduce semantics? Which and how many dimensions will be measured?

Our model will provide an answer to the each one of these questions. What we know about COCOMO [Center(1997)], PSP[Humphrey(2005)], and ONTOCOM [Bontas and Tempich(2005)] is largely based upon empirical studies that investigate: (a) the nominal effort per development, the actual effort, the number of full-time software personnel necessary, the number of instructions per personnel, and the total cost; (b) software size and effort, task and working hours, schedule tracking with earned value, planned value, and earned value; and (c) the man-month effort in hours that are needed for building an ontology.

However, all the studies reviewed so far do not present any metrics for measuring the degree of difficulty of producing semantics with different degrees of complexity.

In our scenario an auditor wants to know how difficult it is to introduce semantics into a company. On the one hand he needs to know the company where the semantics are going to be introduced. On the other hand he needs to know the company that will be introducing the semantics.

The assumptions and limitations of our work are:

A1. There is a company that wants to introduce semantics and its managers agree on it.

A2. There is a company that is willing to introduce semantics and has the resources for achieving it.

A3. The company that will introduce the semantics has a development team with or without semantics infrastructure expertise.

A4. The company that needs the semantics has a problem and is capable of expressing it by means of a tangible resource.

L1. The model measures metrics with some degree of non-accuracy.

L2. The test benchmark is a simulation using the model.

L3. The only answer of the model will be n where n is the distance needed to introduce semantics into a company based on the proposed model.

III. Specifying the Model We Proposed

In this paper we propose a Semantics Difficulty Model (SDM) which is capable of measuring the degree of complexity of introducing semantics into companies. As already explained, we identified three dimensions (see section II) that are compared with our proposed model. SDM works by measuring the difference between each snapshot and the reference model. These measurements are the difficulty level that the company needs to cover in order to introduce semantics. The output is measured by five linguistic values "not at all difficult, slightly difficult, averagely difficult, very difficult and extremely difficult" based on a Euclidian distance [Hartigan and Wong(1979)] algorithm.

Now we will describe each snapshot of the company (company semantic snapshot, data snapshot, and semantic applications snapshot), the maturity level, the reference model, and finally the process for getting the snapshot of the company.

A. Company semantic snapshot

The companies in our study had many activities involving individuals from different cultures, religions, ideals, gender and age with different points of view. Therefore no objective or subjective reality is supported uniformly by multiple domain experts. However, there is still the need to process negotiation and argumentation where the meaning of truth will be aligned and converged. Thus we have the question: how can relevant commonalities and differences in meaning

be captured considering context dependencies of domain experts?

B. Data snapshot

Data can manifest itself in several ways in a company: implicit, explicit, static, dynamic, domain-dependent, and domain-independent.

1) **Implicit.** Implied though not plainly expressed such as images, videos, patents, text messages, audio, national security, documents, call centers, e-mails, on-line forums and customer surveys.
2) **Explicit**. Staged clearly and in detail; are usually structured with metadata as an example: medical record, Twitter, Facebook, corpora, on cloud, websites, blogs, geospatial, reports, catalog formats, Frequently Asked Questions (FAQ)
3) **Static**. Refers to constant data that does not change.
4) **Dynamic** Data consistently generated based on the target function; for example, a transaction is time-dependent.
5) **Domain-dependent** Data are treated as area-specific; for example, the field of medicine.
6) **Domain-independent** No matter what the area is; for example, time, numbers.

C. Metadata snapshot

Metadata is data about data. They are used to know associations with other applications or domains. We identified several levels of metadata and labeled them as follows: systems, dictionaries, taxonomies, ontologies, rules, lexicons, Database management system (DBMS), thesauruses, semantic networks, controlled vocabularies, and schemas.

D. Maturity Level

Based on the levels of Capability Maturity Model Integration $CMMI^{©}$ [University(2000)] we identified six levels for the semantic snapshot of the company: not performed, chaotic, initial, well-defined, quantitatively managed and optimizing. These levels were developed ad hoc and they are defined as follows.

1) **Not performed.** The company does not have any semantics.
2) **Chaotic**. The company has implicit semantics.
3) **Initial.** The company has informal semantics and common business vocabulary that describes multiple views of the truth.
4) **Well-Defined**. The company has semantics in a formal way for humans and information is accessible across the company.
5) **Quantitatively Managed**. Semantics is formal and it is understood by electronic devices and information is used consistently across the company.
6) **Optimizing**. The company is interested in incrementing innovative semantic infrastructure improvements.

E. Reference model

We identified a semantic reference model that splits knowledge into implicit, informal, and formal for electronic devices and humans.

Definition the Semantic Reference Model is a conceptual description relating to the company semantic snapshot as shown in equation (1). It is represented by a core ontology Semantic Company Ontology (SCO) that consist of four disjoint sets C, R, A, and τ where C means concept identifiers (2), R means relation identifiers (3 and 4), A means attribute identifiers (5), and τ means data types (6).

$$SCO := (C, \leq c, R, \gamma_R, \leq_R, A, \gamma_A, \tau) \tag{1}$$

The set C of concepts is:

$$C := \begin{cases} Data, Information, Knowledge, \\ Wisdom, Informal, Implicit, \\ FormalHumans, FormalMachines, \\ Company \end{cases} \tag{2}$$

The set R of relations is:

$$R := \begin{cases} data\_of, information\_of, knowledge\_of, \\ wisdom\_of, data\_in, wisdom\_in, \\ information\_in, knowledge\_in \end{cases} \tag{3}$$

where the relation hierarchy defines that Information has the relation data_in that belongs to Data. Wisdom has the relation Knowledge_in that belongs to Knowledge, following the same logic the rest of the relations are defined as shown in equation (4).

$$\begin{aligned} &\gamma R(data\_in) = (Data, Information) \\ &\gamma R(knowledge\_in) = (Knowledge, Wisdom) \\ &\gamma R(information\_in) = (Information, Knowledge) \\ &\gamma R(wisdom\_in) = (Wisdom, Knowledge) \end{aligned} \tag{4}$$

The set A of attribute identifiers is:

$$A := \begin{cases} report, drawing, photograph, manual, file, \\ productDescription, e-mail, notebook, chart, \\ correspondence, FAQ, document, graph \\ organizationalChart, memo, document, \\ book, bestPractice, specification, taxonomy, \\ conversation, phonecall, metadata, database, \\ real-timeMessages, ontology, program, \\ communicationProtocol, standardProcess, \\ glossary, template, thesaurus, lexicon, \\ catalogFormat, lessonLearned, webservice, \\ UMLdiagram, spreadsheet, topicmap, \\ presentation \end{cases} \tag{5}$$

The set τ of datatypes contains only one element ,string, as shown in equation (6).

$$\tau := (string) \tag{6}$$

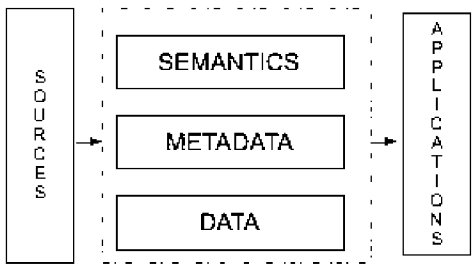

Fig. 1. Process for taking the snapshot of the company

The first axiom defines that the concept Data is equivalent to saying that there is a data which stands in a data_in relation with the corresponding company, following the same logic the rest of the axioms are defined as shown in equation (7).

$$\begin{aligned}
&\forall x(Data(x) \longleftrightarrow \exists y \wedge data_in(x,y) \wedge Company(y)) \\
&\forall x(Information(x) \longleftrightarrow \exists y \wedge information_in(x,y) \\
&\wedge Company(y)) \\
&\forall x(Knowledge(x) \longleftrightarrow \exists y \wedge knowledge_in(x,y) \\
&\wedge Company(y)) \\
&\forall x(Wisdom(x) \longleftrightarrow \exists y \wedge wisdom_in(x,y) \\
&\wedge Company(y))
\end{aligned} \tag{7}$$

F. Process for taking the snapshot of the company

The process for taking the snapshot of the company started when we identified all the sources, the data, the metadata and the semantics involved in order to migrate an application, as shown Fig. 1.

G. SDM cycle

For the SDM cycle four stages are established: simulating data, interpreting data, measurement analysis and calibrating data.

1) **Simulating data** In this step, the first thing that is done is to create a factory of companies. The simulation model provides 2,000 companies that are very useful as input to the SDM.
2) **Interpreting data** The algorithm allows the creation of an initial state based on data, metadata and semantics. The Euclidian distance function is introduced in order to measure the distance between the initial company snapshot and its references models (the company semantic snapshots reference model, data snapshots reference model, and the semantic application snapshots reference model).
3) **Measurement analysis** The size between the snapshots and the references models is conducted by researching the degree of difficulty to introduce semantics into a company. Our results depend on many variables in our efforts to control complexity. The output of SDM will be: not at all difficult, slightly difficult, averagely difficult, very difficult and extremely difficult.

Id	Initial Stage
IE	Information Extraction
DOE	Data and Opinion Extraction
Searching	Searching
Queries	Queries
DDC	Display Diverse Content
QA	Questions Answering
Class	Classification

TABLE I
INITIAL STAGE OF A COMPANY

Id	Final Stage
IData	Integrate data from different sources
RelationDB	New relationships across heterogeneous database
REntities	New relationships between entities
Indexing	Indexing of any content
RUContext	New relationships on the basis of the current users context
SDataI	Establishment of semantic data interchange
OnFlyI	Support on-the-fly information integration
SeparatelyM	Stores data separately from the meaning and content files
MSys	Multiple system

TABLE II
FINAL STAGE OF A COMPANY

4) **Calibrating data** We used a ranking function in order to arrange the results of the iterations. It shows concentrated results based on a set of initial stage transformations.

IV. ANALYSIS

One of the first tasks is to know how the data are used in the companies and what the companies want to do with the data. The first step is to identify where data sources are stored, as shown in Table 1. The identifier is on the left side of the table followed by the possibilities of the initial stage of a company.

Table 2 shows the different semantic applications that are supported by the proposed model SDM.

The following code shows many different combinations to transform one stage to another.

```
IE -> DOE [ label = "S(d)" ];
IE -> Searching [ label = "S(d)" ];
IE -> Queries [ label = "S(d)" ];
IE -> QA [ label = "S(d)" ];
IE -> Class [ label = "S(d)" ];
IE -> IData [ label = "S(d)" ];
IE -> RelationDB [ label = "S(d)" ];
IE -> REntities [ label = "S(d)" ];
IE -> Indexing [ label = "S(d)" ];
IE -> RUContext [ label = "S(d)" ];
IE -> SDataI [ label = "S(d)" ];
IE -> OnFlyI [ label = "S(d)" ];
IE -> SeparateIyM [ label = "S(d)" ];
IE -> MSys [ label = "S(d)" ];
```

```
DOE -> Searching [ label = "S(d)" ];
DOE -> Queries [ label = "S(d)" ];
DOE -> DDC [ label = "S(d)" ];
DOE -> QA [ label = "S(d)" ];
DOE -> Class [ label = "S(d)" ];
DOE -> IData [ label = "S(d)" ];
DOE -> RelationDB [ label = "S(d)" ];
DOE -> REntities [ label = "S(d)" ];
DOE -> Indexing [ label = "S(d)" ];
DOE -> RUContext [ label = "S(d)" ];
DOE -> SDataI [ label = "S(d)" ];
DOE -> OnFlyI [ label = "S(d)" ];
DOE -> SeparateIyM [ label = "S(d)" ];
DOE -> MSys [ label = "S(d)" ];
.................
```

Imagine a possible scenario for introducing semantics. A department wishes to provide a new resource with semantics. The company has a data set and information extraction system (initial stage) that it wishes to transform to integrate data from different sources (final stage) as shown in Fig. 2. Following this step, a interpreting data is employed using

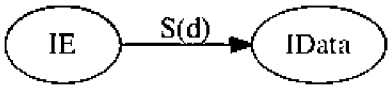

Fig. 2. A Possible scenario for introducing semantics

a population of initial stages. Then the measure function of comparing each stage with the distance from the three proposed dimensions and the Euclidian distance algorithm is dealt with. If the best stage with semantics does not found, the algorithm triggers a new generation; if not, it carries out operations of mutation and crossover until the best solution is found. The conditions for stoppage are: a distance of .3, and a maximum number of 100 iterations.

In this process the distance is measured between the companys semantic snapshot and the reference model proposed. This is done by listing the six best options sorted from best to worst. Finally, the difficulty level is shown as follows: not at all difficult, slightly difficult, averagely difficult, very difficult and extremely difficult.

The model is ready to support 2,000 companies with their corresponding initial and final stages. As a summary, the process for measure the complexity of introducing semantics into a company is shown in algorithm 1.

V. Experimental Results

The experimental results are carried out in two ways: (1) the process to measure the difficulty level to introduce semantics into a company and (2) the best solution of the SDM. From this work, we dimension three arrangements. The first is when the company that wants to introduce semantics does not have any application. So they need to build a complete new application. The second is when the company that wants to introduce the semantics has a legacy application and wants to introduce the semantics in it. We focus our attention on this scenario. The third is when the company that wants to introduce semantics just needs a semantic component in a specific component. For scenarios one and three, the company that will introduce the semantics has three possibilities: to re-use its semantic technology, to build a new semantic technology, and to buy the semantic technology from another company.

Algorithm 1 Process to measure the complexity of introduce semantics into a company

```
1: procedure APB(Data,MetaData,Semantics)
2:     for i ← 1, NumberofData do
3:         for j ← 1, NumberofMetaData(i) do
4:             if StageInitial(j) = SDM(StageInitial) then
5:                 for k ← 1, NumberofElements do
6:                     stageFinal(k) ← data(stageFinal(k));
7:                     stageFinal(k) ← metadata(stageFinal(k));
8:                     stageFinal(k) ← application(stageFinal(k));
9:                     stageFinal(k) ← semantic(stageFinal(k));
10:                end for
11:            end if
12:        end for
13:    end for
14: end procedure
```

A. Process to measure the difficulty level to introduce semantics into a company

One of the experiments achieved is for the second scenario where a company wants to integrate data with a legacy application. The following code shows some of the possibilities to be performed and this is also shown in Fig. 3.

```
InitialStage = [('Stage1','IE'),
          ('Stage2','DOE'),
          ('Stage3','Searching'),
          ('Stage4','Queries'),
          ('Stage5','DDC'),
          ('Stage6','QA'),
          ('Stage7','Class')]

final stage='IData'
```

As a result the SDM simulation model indicates that to transform a search application to integrate data from different sources with semantics is "slightly difficult" in contrast with data and opinion extraction applications that is "very difficult", as shown below.

```
Iteration1   Searching
;SlightlyDifficult
Iteration2   IE ;AveragelyDifficult
Iteration3   DOE ;VeryDifficult
Iteration4   DDivC ;AveragelyDifficult
Iteration5   QA ;AveragelyDifficult
Iteration6   Queries ;VeryDifficult
Iteration7   Classification
;AveragelyDifficult
```

VI. Conclusions

In this paper we have presented a model for answering how difficult it is to introduce semantics into a company. First, we identified three dimensions: a company semantic snapshot, a

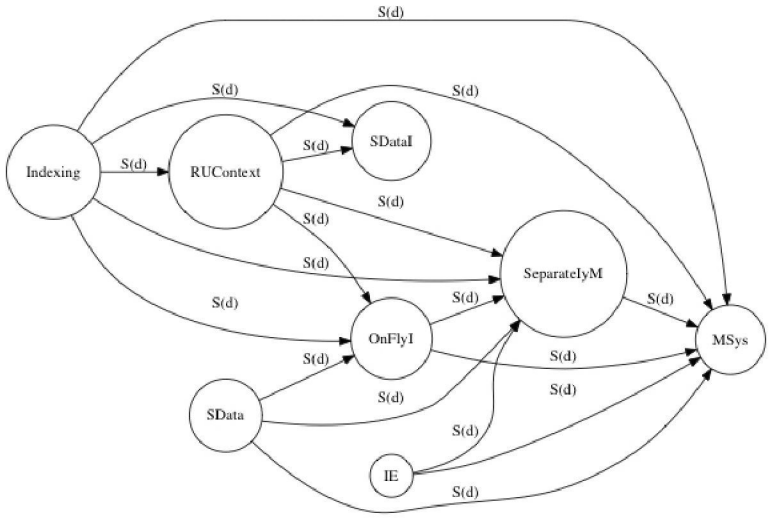

Fig. 3. A partial view of all possible transformations

data snapshot and a semantic application snapshot. Then we recognized the maturity level, the reference model and the process for getting the snapshot of the company. The main results and contributions were that the SDM measured the degree of difficulty to introduce semantics into a company. The most difficult part in our scenario was to know the key factors for introducing semantics into a company.

The semantic technology is mature and has been widely accepted by the major players in the software industry. Even so, the possibility of turning one company into a company with semantics is closer now than seven years ago. Our work in this article serves as a first step towards measuring how difficult it is to introduce semantics into a company. We have shown the degree of semantic complexity with SDM.

SDM helps members of companies to measure the complexity of introducing semantics a priori so they can take early decisions, thus saving time and money, mitigating risks and improving innovation, quality, cost-effectiveness, time to market, and productivity. As each company requires its specific metrics following Paola Di Maio [15] "the metrics, parameters, and evaluation techniques for each project are best set up on an ad hoc basis, by a diverse team, and will have to be calibrated to suit the different aspects of the semantics that is to be emphasized in the project." Finally, further research should be done to test the SDM in real-life companies.

ACKNOWLEDGMENT

The authors would like to thank CONACYT-FC for a grant and OEG of Technical University of Madrid for all the support in this research.

REFERENCES

[Alonso-Ovalle(2006)] Alonso-Ovalle, L.. Disjunction in alternative semantics. Ph.D. thesis; University of Massachusetts at Amherst; Amherst, MA; 2006.

[Nonaka and Takeuchi(1995)] Nonaka, I., Takeuchi, H.. The Knowledge-Creating Company. Oxford University Press; 1995.

[Uschold and Gruninger(1996)] Uschold, M., Gruninger, M.. Ontologies: Principles, methods and applications. Knowledge Engineering Review 1996;11(2):93–155.

[Gruber(1993)] Gruber, T.R.. A translation approach to portable ontology specifications. Knowledge Acquisition 1993;5:199–220.

[Ontoprise(2012)] Ontoprise, . SemanticIntegrator. http://www.ontoprise.de/en/ (Accesed: March 2012); 2012.

[IBM(2012)] IBM, . Watson. http://www-03.ibm.com/innovation/us/watson/index.html (Accesed: March 2012); 2012.

[Oracle(2012)] Oracle, . Bigdata. http://www.oracle.com/technetwork/database/enterprise-edition/overview/index.html (Accesed: March 2012); 2012.

[AG(2012)] AG, . Process intelligence. http://www.softwareag.com/ (Accesed: March 2012); 2012.

[HP(2012)] HP, . Hp labs semantic web research. http://www.hpl.hp.com/semweb/ (Accesed: March 2012); 2012.

[Center(1997)] Center, S.E.. COCOMO II Model Definition Manual. Computer Science Department, University of Southern California, Los Angeles, Ca.; 1997.

[Humphrey(2005)] Humphrey, W.. Psp(sm): a self-improvement process for software engineers. Addison-Wesley Professional; first ed.; 2005. ISBN 9780321579300.

[Bontas and Tempich(2005)] Bontas, E.P., Tempich, C.. How much does it cost? applying ontocom to diligent. Tech. Rep. TR-B-05-20; FU Berlin; 2005.

[Hartigan and Wong(1979)] Hartigan, J.A., Wong, M.A.. A K-means clustering algorithm. Applied Statistics 1979;28:100–108.

[University(2000)] University, C.M.. CMMI for Systems Engineering/Software Engineering, Version 1.02 (CMMI-SE/SW, V1.02) Staged Representation. CMU/SEI-2000-TR-028; 2000.

[Akerkar(2011)] Akerkar, R., editor. Proceedings of the International Conference on Web Intelligence, Mining and Semantics, WIMS 2011, Sogndal, Norway, May 25 - 27, 2011. ACM; 2011. ISBN 978-1-4503-0148-0.